PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : H04L 12/40, 29/06 — A1

(11) International Publication Number: WO 91/15069

(43) International Publication Date: 3 October 1991 (03.10.91)

(21) International Application Number: PCT/US91/01310

(22) International Filing Date: 28 February 1991 (28.02.91)

(30) Priority data:
501,663 29 March 1990 (29.03.90) US

(71) Applicant: SF2 CORPORATION [US/US]; 140 Kifer Court, Sunnyvale, CA 94086 (US).

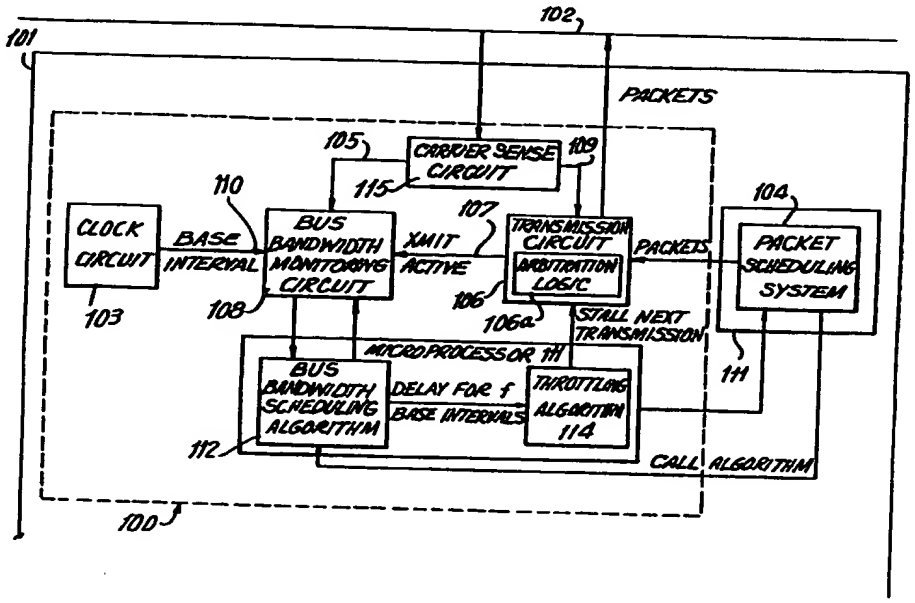
(72) Inventors: JAFFE, David, H. ; 551 South Road, Belmont, CA 94002 (US). JOHNSON, Hoke, S., III ; 16191 Rose Avenue, Monte Sereno, CA 95030 (US). EIDLER, Chris, W. ; 15060 Venetian Way, Morgan Hill, CA 95037 (US).

(74) Agents: ROWLAND, Mark, D. et al.; Fish & Neave, 875 Third Avenue, New York, NY 10022 (US).

(81) Designated States: AT, AT (European patent), AU, BB, BE (European patent), BF (OAPI patent), BG, BJ (OAPI patent), BR, CA, CF (OAPI patent), CG (OAPI patent), CH, CH (European patent), CM (OAPI patent), DE, DE (European patent), DK, DK (European patent), ES, ES (European patent), FI, FR (European patent), GA (OAPI patent), GB, GB (European patent), GR (European patent), HU, IT (European patent), JP, KP, KR, LK, LU, LU (European patent), MC, MG, ML (OAPI patent), MR (OAPI patent), MW, NL, NL (European patent), NO, PL, RO, SD, SE, SE (European patent), SN (OAPI patent), SU, TD (OAPI patent), TG (OAPI patent).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD AND APPARATUS FOR SCHEDULING ACCESS TO A CSMA COMMUNICATION MEDIUM



(57) Abstract

A scheduling mechanism is provided for controlling when the arbitration circuit of a node (10) sharing a CSMA communication medium (14) is to start CSMA arbitration for access to the communication medium once the node has a message ready for transmission, the scheduling mechanism delaying the arbitration circuit (100) from seeking access if total transmission activity (TCU) on the communication medium exceeds a total use threshold (TMU) value and transmission activity (LCS) of the node exceeds a local use threshold value (LMS), and otherwise permitting the arbitration circuit to seek access to the communication medium by arbitration in accordance with a priority value assigned to the node.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	ES	Spain	MG	Madagascar
AU	Australia	FI	Finland	ML	Mali
BB	Barbados	FR	France	MN	Mongolia
BE	Belgium	GA	Gabon	MR	Mauritania
BF	Burkina Faso	GB	United Kingdom	MW	Malawi
BG	Bulgaria	GN	Guinea	NL	Netherlands
BJ	Benin	GR	Greece	NO	Norway
BR	Brazil	HU	Hungary	PL	Poland
CA	Canada	IT	Italy	RO	Romania
CF	Central African Republic	JP	Japan	SD	Sudan
CG	Congo	KP	Democratic People's Republic of Korea	SE	Sweden
CH	Switzerland			SN	Senegal
CI	Côte d'Ivoire	KR	Republic of Korea	SU	Soviet Union
CM	Cameroon	LI	Liechtenstein	TD	Chad
CS	Czechoslovakia	LK	Sri Lanka	TG	Togo
DE	Germany	LU	Luxembourg	US	United States of America
DK	Denmark	MC	Monaco		

METHOD AND APPARATUS FOR SCHEDULING ACCESS TO A CSMA COMMUNICATION MEDIUM

Background of the Invention

The present invention relates to carrier sense multiple access (CSMA) protocols. In particular, the present invention relates to an improvement in scheduling transmissions within a CSMA protocol having a prioritized set of statically assigned time slots.

CSMA protocols generally can be considered as methods for distributing management of a communication medium among users of the medium. The medium with which CSMA protocols are concerned is a single-carrier communication medium, such as a co-axial communication bus in a computer network or system or a transmission channel in a satellite network, in which an arbitration mechanism must be embedded in the same carrier used to convey data.

Many variants of CSMA protocols are known. In some of these variants, prioritized arbitration is accomplished by assigning each node sharing the communication medium a unique time slot during which the node may initiate transmission if no other node having a higher priority (i.e., an earlier time slot) has already begun a transmission. The set of time slots for all nodes is synchronized to within one propagation time interval with the trailing edge of the

carrier signal that follows the end of a transmission on the communications medium.

Where time slots are assigned statically (e.g., the time delay for each node of a network is set at the time of installation and is not changed thereafter), nodes having low priority may experience long delays in obtaining access to the communication medium, particularly during periods of heavy traffic.

To place an upper boundary on transmission delays for all messages originating at any node, "fair" CSMA protocols have been developed in which assigned priorities are varied such that all nodes awaiting access at a given instant are allowed to access the communication medium once before any node is given a second chance. For a discussion of such a "fair" protocol, see "BRAM: The Broadcast Recognizing Access Method," Chlamtac, I., Franta, W.R., and Levin, K.D., IEEE Transactions On Communications, Vol.Com-27, No.8, August 1979, pp. 1183-1190; and "Message-based priority access to local networks," Chlamtac, I. and Franta, W.R., Computer Communications, Vol. 3, No. 2, April 1980, pp. 77-84.

In at least certain network configurations, such as a computer network including one or more server nodes, a fair CSMA protocol does not necessarily provide good system throughput. This is because an application may require that a server node have high priority access to the communication medium to handle a large number of data requests.

For example, where a server node comprises a mass storage subsystem that handles the mass storage needs of several other nodes (referred to hereafter as "served nodes"), the server node is likely to be a bottleneck in the network if it is limited to seeking access to the network bus on a fair basis with each of

the served nodes. In general, to avoid a bottleneck situation, the server node should be allowed a share of the bus bandwidth at least equal to the sum of the bandwidth shares of the served nodes (assuming that the mass storage subsystem is capable of operating at such a rate). In a prioritized CSMA protocol, this level of system performance (i.e., share of bus bandwidth) can only be guaranteed if the server node is statically assigned the highest priority for bus access. Otherwise, the performance level of the subsystem, as well as that of the entire system, may be limited to less than its potential capability. Nodes having priority higher than that of the mass storage subsystem node may capture a share of bus bandwidth that does not leave sufficient bandwidth available for the mass storage subsystem to perform up to its potential maximum throughput capacity.

On the other hand, a server node having statically-assigned high priority access also may, in some applications, cause system throughput to be degraded because the server node uses a large share of the communication medium bandwidth at a time when other nodes need to transmit. System throughput in such a case would be improved if the share of bandwidth used by the server node could be dynamically adjusted based on attributes of particular applications.

It may also occur in a network that a server node is capable of transmitting data at a faster rate than a receiving node can successfully receive the data. System throughput suffers if, as a result of such difference in data handling capacities among nodes, a serving node must re-transmit data that a slower node was incapable of receiving successfully. This problem can be avoided by providing a means for

the server node to lower its effective transmission rate when transmitting to slower nodes.

It would thus be desirable within a CSMA protocol to permit a node to dynamically adjust its share of transmission time on the communication medium to improve system throughput in different network applications.

It would further be desirable to be able to implement such an adjustable control mechanism in a manner that would allow a node to vary its share of the communication medium without violating priority arbitration schemes incorporated in CSMA protocols, such that the adjustable control mechanism could be utilized by a device in networks having various types of CSMA protocols.

It would also be desirable to be able to provide a mechanism for selectively reducing the transmission rate of transmissions to nodes having low data handling capacity.

Summary of the Invention

It is an object of the present invention to provide within certain types of prioritized CSMA protocols a mechanism for allocating an adjustable share of transmission bandwidth to particular nodes.

It is a further object of the present invention to provide an arbitration circuit incorporating a configurable scheduling mechanism which can be implemented with any CSMA protocol of the type having at least one set of prioritized arbitration time slots.

It is an additional object of the present invention to provide a mechanism for determining a receiving node's data transfer rate and for generating

for each receiving node a scaled speed estimate based on this determination.

It is also an object of the present invention to provide a mechanism for scheduling a delay between transmissions to a node to reduce the rate at which data is transmitted to a slower receiving node. The term "per-node throttling" is used herein to refer to the function of such a mechanism.

In accordance with the present invention, a dynamic scheduling mechanism having configurable parameters is provided for controlling when the arbitration circuit of a node sharing a CSMA communication medium is to seek access to the communication medium once the node has a message ready for transmission, the scheduling mechanism delaying the arbitration circuit from seeking access if (1) total transmission activity on the communication medium exceeds a total use threshold value and (2) transmission activity of the node exceeds a local use threshold value, and otherwise permitting the arbitration circuit to seek access to the communication medium by arbitration in accordance with a priority value assigned to the node.

There is further provided a dynamic per-node throttling mechanism for delaying sequential transmissions from a transmitting node to any receiving node in the network having a slower data transfer rate than the transmitting node. The relative speed of a receiving node is determined by the frequency with which negative acknowledgment responses have been received from that node in the past. An algorithm is provided for introducing dynamically variable and node-specific delay periods between sequential transmissions to slower nodes based on the determined relative speed of each individual receiving node.

Brief Description of the Drawings

The above and other objects and advantages of the present invention will be apparent upon consideration of the following detailed description, taken in conjunction with the accompanying drawings, in which like reference characters refer to like parts throughout, and in which:

FIGS. 1A and 1B illustrate, respectively, a computer network having a plurality of nodes, and a schematic diagram of an arbitration system for a node on a CSMA protocol communication bus including scheduling mechanisms in accordance with the present invention;

FIG. 2 is a diagram of a gauge of the bandwidth of a shared communication medium, such as the communication bus of FIG. 1;

FIG. 3 is a schematic diagram of an embodiment of a circuit for use in implementing the configurable bandwidth monitoring mechanism of FIG. 1;

FIG. 4 is a diagram of a flow chart showing the steps of an exemplary implementation of the method of the configurable bandwidth scheduling mechanism of the present invention;

FIGS. 5A-5C are diagrams of a flow chart showing the steps of an exemplary implementation of the method of the per-node throttling mechanism of the present invention; and

FIG. 6 is a diagram of a table reflecting various values and parameters for the per-node throttling mechanism of the present invention.

Detailed Description Of The Invention

Referring to FIG. 1A, the present invention particularly concerns prioritized CSMA protocols in which a set of time slots are statically assigned to

the nodes 10 of a network 12 sharing a common communication medium 14. Each slot of the set may be assigned to one and only one node, or may be assigned to a plurality of nodes. A node may even be assigned or assign itself to more than one time slot in the set if it has a particularly great need for time on the medium. The set of time slots provides priority access control for nodes wishing to transmit when the communication medium is busy. This is typically accomplished in conventional collision avoidance type CSMA protocols as follows.

Each transmission on the communication medium is accompanied by a carrier signal. At the end of a transmission on the medium, the carrier signal ceases and the medium enters an idle state for a fixed period of time. During this fixed period of time, only a node responding to the previous message is allowed to transmit. Thus interference with a required response message (e.g., an ACK or NAK message) is prevented. Following the fixed delay period, a time slot period begins. During this period each node has one or more time slots in which it may capture the communication medium without interference from other nodes, thus establishing a priority system for capture of the communication medium. A node wishing to transmit a message waits only until its time slot occurs, at which time it can freely access the communication medium without interference if the medium is idle. The time slots assigned to nodes in the network depend on the priority the nodes are to be given -- the higher the priority, the closer the time slot is to the beginning of the time slot period.

In a network in which such a prioritized CSMA protocol is used to control access to a common communication medium, each node generally includes an

arbitration circuit that monitors transmission activity on the communication medium and determines when the node is allowed to access the medium to transmit a message. This determination is made in accordance with the rules of the CSMA protocol, and is typically accomplished using hardware logic and associated circuitry which has been programmed according to the prioritizing function of the CSMA protocol.

FIG. 1B illustrates a configurable arbitration circuit 100 in accordance with the principles of the present invention. Arbitration circuit 100 is shown as part of a node 101 in a local area network or a distributed computing system, and is connected to a communication bus 102 of the network or system. Communication bus 102 may be realized using any one of various known technologies such as fiber optics, coaxial cable or microwave channel. Bus 102 may comprise a single bit-serial line. In a preferred embodiment, arbitration circuit 100 is implemented as part of a bus interface unit for connecting a node to a serial communication path of a packet-switching type network.

Messages to be transmitted on bus 102 are provided to arbitration circuit 100 by a packet scheduling system 104 that assembles the messages into packets. Packet scheduling system 104 is controlled by a processor 111. The packets are provided to transmission circuit 106, which includes arbitration logic circuitry 106a for determining, in accordance with a lower level of the CSMA protocol, when node 101 may take control of bus 102 to initiate a transmission. Carrier sense circuitry 115 detects the state of bus 102 (busy or idle) by detecting the presence or absence of a carrier signal on the bus and generates carrier detect signals 105 and 109. The arbitration logic

circuitry 106a, in response to transmission circuit 106 receiving a packet from packet scheduling system 104 ready for transmission, checks the carrier sense circuitry 115 to determine if a carrier is present on bus 102. If a carrier is present, signifying that another node sharing bus 102 has control of the bus, the arbitration logic circuitry 106a waits for the carrier sense circuitry 115 to detect the falling edge of the carrier signal.

When a falling edge is detected, the arbitration logic circuitry 106a enters a first timed waiting state. This waiting state allows a receiving node in the network time to send an acknowledgement message in response to the termination of the last transmitted message without having to arbitrate for control of bus 102. The duration of this waiting state is typically equal to the amount of time required for a signal to propagate from one end of bus 102 to the other, plus the amount of time required to detect a carrier. The waiting state also allows a node that has access to bus 102 to continue to use it in a series of transmissions.

When the first timed waiting period elapses, the arbitration logic circuitry 106a then enters a second timed waiting state during which it checks the carrier sense circuitry 115 for the presence of a carrier signal on bus 102. The duration of the second waiting period is determined by the priority value assigned to the node. The node must wait this second period to allow nodes having higher priority values (and thus shorter waiting periods) an opportunity to initiate a transmission. The nodes are each given one or more limited time slots in which to initiate transmission, the duration of each time slot typically being the same as that of the first waiting period.

If bus 102 becomes active during the second waiting period (signifying that a node of higher priority has won control of the bus), the arbitration logic circuitry 106a again awaits the end of the transmission and repeats the first and second timed waiting periods. If bus 102 remains idle throughout the second waiting period, the arbitration logic circuitry 106a may win control of the bus by initiating a transmission during its assigned time slot. Transmission circuit 106 preferably includes programmable clock and counter circuits for measuring the first and second waiting periods. These circuits are programmed in accordance with the physical properties of the network (e.g., distance between nodes, number of nodes, etc.) and the assigned priority value of node 101. The counter circuit, for example, may be a modulo counter that is programmed to repeatedly count down the necessary waiting periods. The counter is reset each time a falling carrier edge is detected. Thus, whenever the node has a packet ready for transmission, the node waits until the modulo counter next reaches its terminal count (e.g., zero) before initiating a transmission.

In conventional modes of network communication, successful receipt of a message is normally acknowledged by the receiving node immediately after the message transmission is completed. This acknowledgement typically takes the form of an acknowledgement message (ACK) which is transmitted by the receiving node to the transmitting node during a fixed period of time following the end of the previous transmission. A negative acknowledgement response (NAK) is transmitted by the receiving node to the transmitting node instead if the message was successfully transmitted to the intended receiving

node, but that node was too busy to properly buffer the message. If a proper acknowledgement (ACK or NAK) is not received within this fixed period (e.g., because the intended receiving node did not receive the transmission), the transmitting node considers the transmission to have failed.

The transmitting node, upon receiving either no acknowledging response, or an acknowledgement response indicating that the message packet was properly received but could not be handled by the receiving node (e.g., a NAK), will attempt to re-transmit the message. This acknowledgement process and the rescheduling of transmitted messages is accomplished by the packet scheduling system 104 of the node.

When transmitting a message, transmission circuit 106 provides a XMIT ACTIVE signal 107 to bus bandwidth monitoring circuit 108 indicating that the node is transmitting. Bus bandwidth monitoring circuit 108 receives from carrier sense circuitry 115 a carrier detect signal 105 indicating the presence or absence of transmission activity on bus 102. Bus bandwidth monitoring circuit 108 and transmission circuit 106 may use the same carrier sense circuitry 115 as shown in FIG. 1B.

The function of bus bandwidth monitoring circuit 108 is to generate (1) a measure of the amount of the bandwidth of bus 102 that is used by all nodes sharing the bus (total bus bandwidth use), and (2) a measure of the amount of bandwidth used by the node 101 which includes arbitration circuit 100 (local bus bandwidth use). The first measure is taken by sampling bus 102 at a constant rate for the presence of a carrier signal. The second measure is taken by

sampling XMIT ACTIVE signal 107, preferably at the same constant rate.

A synchronous base interval signal 110 is provided by clock circuit 103 to bus bandwidth monitoring circuit 108 to clock the sampling functions. The sampling rate can be chosen as desired. For example, a sampling rate may be defined such that the time between samples is equal to the period of a time slot assigned to a node sharing bus 102 (which is based on propagation delay across bus 102 and the amount of time required to detect a carrier). The measures of total bus bandwidth use and local bus bandwidth use generated by bus bandwidth monitoring circuit 108 are provided to bus bandwidth scheduling algorithm 112. Bus bandwidth scheduling algorithm 112, which is preferably implemented using a subroutine executed by processor 111, controls the rate at which transmission circuit 106 arbitrates for control of bus 102 in response to the scheduling of packets by packet scheduling system 104. Bus bandwidth scheduling algorithm 112 may also be implemented using other types of conventional logic circuitry, such as a programmable state machine logic circuit.

Scheduling algorithm 112 may exercise control over transmission circuit 106 in any one of several ways. For example, scheduling algorithm 112 may be capable of temporarily disabling the transmission circuit 106 to prevent it from transmitting a packet scheduled by packet scheduling system 104, or it may be capable of temporarily preventing packets from being provided to transmission circuit 106 by package scheduling system 104.

Bus bandwidth scheduling algorithm 112 operates to reduce the rate at which node 101 transmits packets when total activity on the bus exceeds a first

threshold percentage of the available bandwidth of bus 102 and the share of bus bandwidth used by node 101 exceeds a second threshold percentage of the available bandwidth of bus 102. Pertinent values and parameters are labeled in FIG. 2, which illustrates a bandwidth gauge 200 of bus 102.

Referring to FIG. 2, total transmission activity on bus 102 is indicated by Total Current Use (TCU). The above-stated first threshold percentage of available bandwidth is indicated by Total Maximum Use (TMU). The share of available bus bandwidth used by node 101 for transmissions is indicated by Local Current Share (LCS), and the above-stated second threshold percentage is indicated by Local Minimum Share (LMS).

Using these labels, the function of bus scheduling algorithm 112 can be stated as follows: if total transmission activity on bus 102 (TCU) exceeds a maximum threshold percentage (TMU) of available bandwidth on bus 102, which may, for example, be 80% (i.e., the bus is busy 80% of the time), reflecting that overall demand for use of the bus is high, and if node 101 has a share of transmission activity on the bus (LCS) exceeding a minimum threshold value (LMS), then bus bandwidth scheduling algorithm 112 introduces a delay into the rate at which node 101 arbitrates for access to the bus to reduce the total transmission activity (TCU) value to (or below) the maximum threshold percentage (TMU) and/or to reduce the share of node 101 (LCS) to (or below) the minimum threshold (LMS).

The threshold values of Total Maximum Use (TMU) and Local Minimum Share (LMS) are dependent on the particular configuration of the network, and the particular application being run on the network. These

threshold values can be varied from one network to another, and from one application to another, as appropriate to achieve optimum system throughput. For any particular configuration and application the appropriate threshold values can be determined empirically.

The operation of an exemplary embodiment of bus bandwidth scheduling algorithm 112 is illustrated by FIGS. 3 and 4. FIG. 3 shows a circuit 300 incorporated in bus bandwidth monitoring circuit 108 for generating measures of total and local bus bandwidth use. Circuit 300 includes AND logic gates 302 and 304 which respectively sample carrier detect signal 105 and XMIT ACTIVE signal 107 in response to each clock cycle of synchronous base interval signal 110. The output of AND gate 302 is coupled to the clock input of accumulator 306, which is incremented once for each time a carrier signal is detected when sampled by AND gate 302.

Likewise, the output of AND gate 304 is coupled to the clock input of accumulator 308, which is incremented once for each time node 101 is transmitting when sampled by AND gate 304. Accumulators 306 and 308 are reset at a regular interval defined by a count value stored in interval register 310 by scheduling algorithm 112. This interval count value is, in turn, loaded into counter circuit 312, which decrements the value once for each clock cycle of base interval signal 110. When counter circuit 312 reaches zero, accumulators 306 and 308 are loaded respectively into latch registers 314 and 316, and the accumulators are reset to begin a new count.

The values stored in latch registers 314 and 316 provide scheduling algorithm 112 with respective measures of total transmission activity on bus 102 and

local transmission activity of node 101 over an interval defined by the scheduling algorithm. The values are updated every interval, such that the measures are kept current. In performing its control function, the algorithm uses the value of latch register 314 as the previously described parameter Total Current Use (TCU), and the value of latch register 316 as the previously described parameter Local Current Share (LCS). A flow chart of an exemplary subroutine for implementing the control function of scheduling algorithm 112 using processor 111 is shown in FIG. 4.

Referring to FIG. 4, the subroutine is activated by processor 111 when packet scheduling system 104 has a packet ready for transmission (step 400). As part of packet scheduling system 104, processor 111 may maintain a queue of identifiers corresponding to ready packets. If a packet is ready, as indicated by an identifier corresponding to the packet being placed in the queue, processor 111 reads the value of latch register 314 and compares the value to the programmed threshold value Total Maximum Use (TMU). If the Total Current Use (TCU) value of latch register 314 is less than or equal to the Total Maximum Use threshold value (TMU), processor 111 enables transmission circuit 106 to arbitrate for control of bus 102 to transmit the ready packet (or, depending on the implementation, allows the ready packet to be given to transmission circuit 106 by packet scheduling system 104). Steps 402 and 408.

Otherwise, processor 111 reaches another decision point (step 404) in which it reads the value of latch register 316 and compares the value to the programmed Local Minimum Share threshold value (LMS). If the Local Current Share value (LCS) of latch

register 316 is less than or equal to the Local Minimum Share threshold value (LMS), the algorithm allows the ready packet to be transmitted. Otherwise, the algorithm delays the packet (step 406) for a period of time (e.g., a certain number of cycles of base interval signal 110) before allowing the ready packet to be transmitted.

The delay period is defined by a function (f), which preferably provides a delay value that reduces node 101 bus bandwidth use such that either (1) the Local Current Share (LCS) value is reduced to (or below) the Local Minimum Share (LMS) threshold value, or (2) the Total Current Use (TCU) value is reduced below the Total Maximum Use (TMU) threshold value. This function may be implemented in various ways. For example, function (f) may comprise an indexing function for pointing to a particular delay value in a table of various empirically determined delay values. Alternatively, function (f) may comprise a formula for calculating the delay value, such as the following:

f(TCU, TMU, LCS, LMS) = the lesser of (LCS-LMS) and (TCU-TMU) [in cycles of base interval signal 110]

Scheduling algorithm 112 may alternatively calculate a measure of free time on bus 102 (e.g., by subtracting the Total Current Use value (TCU) of latch register 314 from the value of interval register 310), and use the value of bus free time as a parameter (instead of using its complement).

As can be seen from the above-described embodiments of bus bandwidth scheduling algorithm 112, the control function of the algorithm can be varied (e.g., by varying the threshold parameters) to vary the

share of bus bandwidth controlled by arbitration circuit 100.

As shown in FIG. 1B, the present invention further includes a per-node throttling algorithm 114 executed by processor 111 which restrains (or throttles) the rate at which node 101 may transmit to any particular receiving node sharing bus 102. Repeated high speed transmissions to a single node may result in packets being received by the node at a greater rate than they can be buffered and processed, such that the packets are discarded by the receiving node and must be retransmitted.

Retransmission of a packet wastes bandwidth on bus 102. This wasted bandwidth can be reduced by requiring node 101 to limit the rate at which it transmits packets to a slower receiving node. Throttling algorithm 114 accomplishes this limiting function by introducing a delay between sequential transmissions to the slower receiving node. More particularly, throttling algorithm 114 monitors NAKs received by node 101 and computes, for each receiving node in the network, a NAK rate per packet transmitted by node 101. Throttling algorithm 114 requires that a certain minimum delay period elapse between transmissions to a particular node if the past performance of that node reflects a per-packet NAK rate that exceeds a threshold value. The threshold value can be varied from network to network, from application to application and from node to node as appropriate to optimize system throughput in each case.

Preferably, where a minimum delay is imposed on transmissions to a particular node, the duration of the minimum delay is a computed value based on the observed per-packet NAK rate for that node. Although the data handling capabilities of particular nodes may

be known prior to the beginning of an application, these capabilities will normally vary during operation, such that optimum system performance is more likely to be achieved by detecting such variations "on the fly" and adjusting the length of any minimum delay imposed to take into account the magnitude of the variations. For this purpose, throttling algorithm 114 preferably uses the per-packet NAK rate as a scaled estimate of the current data handling capability, or speed, of each receiving node, not only for determining whether or not to delay transmissions to the receiving node, but also to determine the duration of the delay.

An example of an embodiment of throttling algorithm 114 is described below. For purposes of illustration, the exemplary embodiment of throttling algorithm 114 is described in the context of a packet-switching type network in which node 101 is a server node for a plurality of other nodes (served nodes). A flow chart of the algorithm is shown in FIGS. 5A-5C.

The steps of throttling algorithm 114 shown in FIGS. 5A-5C are preferably implemented using subroutines executed by processor 111. First, during initialization of the server node, a subroutine (500 of FIG. 5A) is called to create a data structure in the memory of processor 111. An illustrative diagram of such a data structure 600 is shown in FIG. 6. The data structure includes columns for entering the following information for each served node on the network with which node 101 must communicate: CURRENT NAK RATE, DELAY INTERVAL, TOTAL NAKS, TOTAL PACKETS, REMOTE PORT TYPE DELAY INTERVAL and TIME OF LAST PACKET.

Each row of the data structure represents information for a different served node. The rows are identified by numbers assigned to the served nodes by processor 111. Step 502 of FIG. 5A. The served node

numbers represent the offset in memory of each corresponding row from the top of data structure 600. Processor 111 addresses the various rows of the data structure by combining the respective served node number with the beginning memory address of the data structure in address pointer 602. Based on parameter values input by the operator or some other source at initialization, processor 111 enters into the REMOTE PORT TYPE DELAY INTERVAL column a delay value for each served node. Step 504. Processor 111 also enters this value into the DELAY VALUE column as an initial value. Step 506.

Preferably, when the server node is connected, the node exchanges information with served nodes concerning the type of ports those served nodes are using, and the operator chooses initial delay values for the network nodes depending on known limitations in the data transfer capacities of the port types for those nodes. For some or all nodes, the operator may choose to begin with no initial delay. The operator may also choose to set the same initial delay for transmissions to all nodes.

During run time, the delay values in the DELAY INTERVAL column are used by processor 111 to set the minimum delay between transmissions to each node. Processor 111 maintains a queue in which it identifies packets that are ready for transmission and the number of the served node to receive the packets. When scheduling algorithm 112 has determined that node 101 may transmit (step 408), scheduling algorithm 112 calls subroutine 508 of throttling algorithm 114. Subroutine 508 checks the entry in the DELAY INTERVAL column corresponding to the served node that is to receive the next packet and determines whether that value (if non-zero) is greater than the time that has elapsed since

the last packet was sent to that served node. Step 510. The time of the last transmission is found in the TIME OF LAST PACKET column of data structure 600.

If the time since the last transmission to that served node is equal to or exceeds the value of DELAY INTERVAL, throttling algorithm 114 permits the transmission to take place without further delay. Step 512. Otherwise, processor 111 delays the transmission until the delay interval is reached. Step 514.

After the transmission of a packet, processor 111 updates the value of the entries in the TOTAL PACKETS and TIME OF LAST PACKET columns for the served node that was sent the packet. Step 516. The entry reflects the total number of packets sent to a particular served node. Processor 111 then awaits a response from the served node. If the served node acknowledges successful receipt, the processor exits the subroutine. Step 518. If a NAK is received from the served node, processor 111 updates the value of the entry in the TOTAL NAKS column for that served node. Step 520. Processor 111 updates the TOTAL NAKS column for each NAK received. Processor 111 also calculates a CURRENT NAK RATE value for the served node by dividing the value of the entry in the TOTAL NAKS column by the value of the entry in the TOTAL PACKETS column, and determines whether the value of DELAY INTERVAL should be changed. Steps 522 and 524.

Processor 111 preferably increases the value of DELAY INTERVAL for a served node if the value of CURRENT NAK RATE for that served node increases. This may be accomplished in various ways. The particular method chosen is implementation specific. As one example, a look-up table may be provided according to which different DELAY INTERVAL values are specified for each served node for different values, or ranges of

values, of CURRENT NAK RATE. Each time the CURRENT NAK RATE for a served node increases, processor 111 checks the look-up table to determine if the new value corresponds to a new DELAY INTERVAL value. In this manner, it may be provided that the DELAY INTERVAL value remains at zero until an initial non-zero threshold value for CURRENT NAK RATE is exceeded. Alternatively, processor 111 may increment the DELAY INTERVAL value by a fixed number each time the CURRENT NAK RATE increases, or another function may be defined for calculating the value of DELAY INTERVAL based on the value of CURRENT NAK RATE.

Likewise, the value of DELAY INTERVAL for a particular node may be decreased to reflect downward changes in the CURRENT NAK RATE value. Processor 111 may also check at this time, or on a periodic basis, for abnormally high values of CURRENT NAK RATE indicative of a malfunctioning remote port.

It is a known practice in network communication for a server node to schedule transmissions to various receiving nodes in an interleaved manner to provide a balanced throughput. Throttling algorithm 114 may take advantage of this technique while awaiting a delay interval for a particular node to elapse by transmitting packets to other nodes during the interval. FIG. 5C illustrates a subroutine 526 that may be called by processor 111 to accomplish this.

Processor 111 maintains a modulo-n served node queue counter, n being equal to the number of nodes served by the mass storage subsystem. Each served node is represented by a different value of the counter. These values correspond to the served node numbers that identify the rows of data structure 600. When a packet ready for transmission by node 101 is

delayed by throttling algorithm 114, the value of the counter is set equal to the next modulo-n value following the number of the served node that is to receive the delayed packet. Step 528. Processor 111 checks a queue of other packets ready for transmission to determine if any are destined for the served node indicated by the modulo-n counter. Step 530. Such a queue is preferably maintained for each served node. If a packet is ready, processor 111 then determines whether for that particular served node there is a need to delay transmission of the ready packet to prevent the bandwidth capacity of the served node from being exceeded. Step 532. If not, processor 111 causes the packet to be transmitted, and determines if another packet is ready for transmission to that served node (thus allowing consecutive packets to be transmitted to a particularly fast port). Step 534. If no packet is ready for transmission to the served node the count is incremented and the queue for the next served node is checked for a ready packet. This process repeats until the delay interval of the initial packet times out.

Thus a novel method and apparatus for scheduling transmissions within a CSMA protocol communication medium have been described. One skilled in the art will appreciate that the present invention can be practiced by other than the described embodiments, and in particular may be incorporated in circuits other than the described mass storage subsystem. The described embodiment is presented for purposes of illustration and not of limitation, and the present invention is limited only by the claims which follow.

WHAT IS CLAIMED IS:

1. In a network of a plurality of nodes sharing a common communication medium, the network having a CSMA protocol including at least one set of prioritized arbitration time slots assigned to the plurality of nodes sharing the communication medium, a first node having transmission means for accessing the communication medium, the transmission means comprising:

means for arbitrating for access to the communication medium in accordance with a priority value assigned to the first node; and

means for controlling when the transmission means is to seek access to the communication medium, the controlling means delaying the transmission means from seeking access if total transmission activity on the communication medium exceeds a total use threshold value and transmission activity of the first node exceeds a local use threshold value, and otherwise permitting the transmission means to seek access to the communication medium by arbitration in accordance with the priority value assigned to the first node.

2. The node of claim 1, further comprising means for delaying a transmission to a second node for at least a minimum delay period after a prior transmission from the first node to the second node, the minimum delay period being adjustable in accordance with changes in the data handling capability of the second node.

3. The node of claim 2, further comprising means for interleaving a transmission from the first node to a third node during the minimum delay period

between sequential transmissions from the first node to a second node.

4. The node of claim 1, wherein the controlling means comprises:

means for determining a measure of current local use of the communication medium by the node;

means for determining a measure of current total use of the communication medium by the plurality of nodes sharing the communication medium;

means for temporarily disabling the transmission means when the determined measure of current total use is greater than the total use threshold value and the determined measure of current local use is greater than a local use threshold value.

5. The node of claim 4, wherein the disabling means includes means independent of the arbitration protocol for computing a delay period which varies as a function of the determined measures of current total and local use and the total and local use threshold values.

6. The node of claim 1, wherein transmissions on the communication medium include messages assembled into packets, and the node further comprises a packet scheduling system which assembles a message ready for transmission on the communication medium into a packet and provides the packet to the transmission means, and wherein the controlling means comprises:

means for determining a measure of local use of the communication medium by the node;

means for determining a measure of total use of the communication medium by the plurality of nodes sharing the communication medium;

means for temporarily preventing the packet scheduling system from providing a packet to the transmission means when the determined measure of total use is greater than the total use threshold value and the determined measure of local use is greater than a local use threshold value.

7. The node of claim 6, wherein the preventing means includes means independent of the arbitration protocol for computing a delay period which varies as a function of the determined measures of current total and local use and the total and local use threshold values.

8. A method for managing use by a local node of a carrier sense multiple access communication medium, which medium is shared by a plurality of nodes, the method comprising the steps of:

determining a measure of past local use of the communication medium by the local node;

determining a measure of past total use of the communication medium by the plurality of nodes sharing the communication medium; and

adjustably scheduling a transmission by the local node over the communication medium as a function of the determined measures of past local and total use of the communication medium, wherein the scheduling function delays the transmission if the determined measure of past total use is greater than a total use maximum threshold value and the determined measure of past local use is greater than a local use minimum threshold value.

9. The method of claim 8, wherein the measure of past local use is determined by:

generating a signal indicative of whether the local node is transmitting;

sampling the generated signal at a predetermined rate; and

counting over a predetermined interval the number of times the local node is transmitting when the generated signal is sampled.

10. The method of claim 8, wherein the measure of past total use is determined by sampling the communication medium at a predetermined rate and counting over a predetermined interval the number of times the communication medium is busy when sampled.

11. A method for controlling the share of available bandwidth on a CSMA communication medium used for transmission by a device demanding control of the medium, the method comprising the steps of:

sensing the communication medium to determine over a period of time a measure of how often a carrier signal is present on the communication medium;

prescribing a minimum bandwidth share value for the device;

limiting the share of available bandwidth used for transmission by the device to the prescribed minimum bandwidth share value when the measure of carrier signal presence exceeds a threshold value.

12. A system for controlling access by a node to a multi-node CSMA communication medium, the system comprising:

means for monitoring transmission activity on the communication medium; and

means for controlling the bandwidth share used by the node as a function of the monitored total transmission activity, wherein the controlling means defines a programmable limit on the share of bandwidth used by the node when free time on the communication medium is less than a programmable threshold value.

13. The system of claim 12, wherein the monitoring means comprises:

first sampling circuitry for providing an indication of the presence or absence of a carrier signal on the communication medium at particular intervals;

first counting circuitry responsive to the first sampling circuitry for determining the number of intervals at which the presence of a carrier signal is indicated over a selected period of time;

second sampling circuitry for providing an indication of a transmission on the communication medium by the node at the particular intervals; and

second counting circuitry responsive to the second sampling circuitry for determining the number of intervals at which a transmission by the node is indicated over the selected period of time.

14. The system of claim 12, wherein the controlling means comprises:

programmable means for selectively scheduling a delay of variable length in accessing the communication medium when the node seeks to transmit.

15. The system of claim 13, wherein the controlling means comprises:

programmable means for selectively scheduling a delay of variable length in accessing the communication medium when the node seeks to transmit.

16. The system of claim 15, wherein the programmable means derives a measure of the free time on the communication medium from the first counting circuitry and derives a measure of the share of bandwidth used for transmission by the node, and wherein the programmable means schedules the delay of variable length if the measure of free time is less than a first programmed threshold value and the measure of bandwidth share of the node exceeds a second programmed threshold value.

17. The system of claim 16, wherein the programmable means includes a processor, and wherein the length of the delay is computed by the processor such that the share of bandwidth used for transmission by the node is reduced at least to the second programmed threshold value.

18. In a network including a plurality of nodes, a system for limiting the rate at which a first network node transmits sequential transmissions to a second node in the network, the system comprising:

means for determining an estimate of the data handling capability of the second node from past transmissions between the first and second nodes;

means for defining a minimum delay period for sequential transmissions to the second node as a function of the determined estimate of the data handling capability of the second node; and

means for delaying a scheduled transmission to the second node if the minimum delay period has not elapsed since a prior transmission from the first node to the second node.

19. The system of claim 18, wherein the means for determining an estimate of the data handling capability of the second node comprises means for detecting over an interval the number of negative acknowledgment responses received from the second node in response to attempted transmissions from the first node to the second node.

20. The system of claim 18, further comprising means for interleaving transmissions to different nodes in the network during the delay of a transmission to the second node by the delaying means.

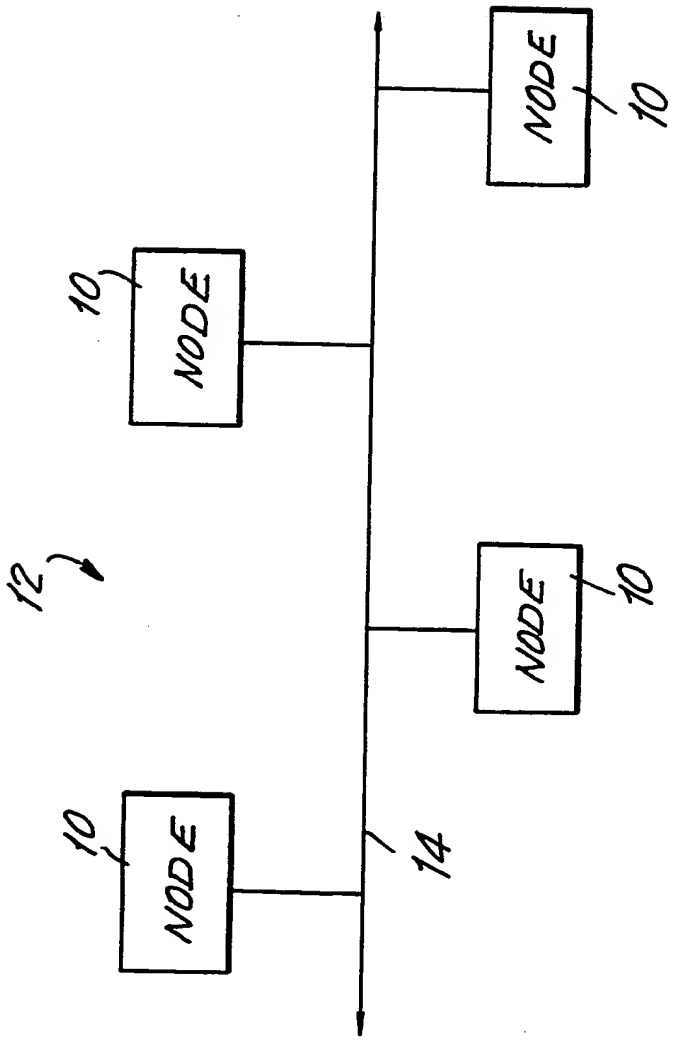


FIG. 1A

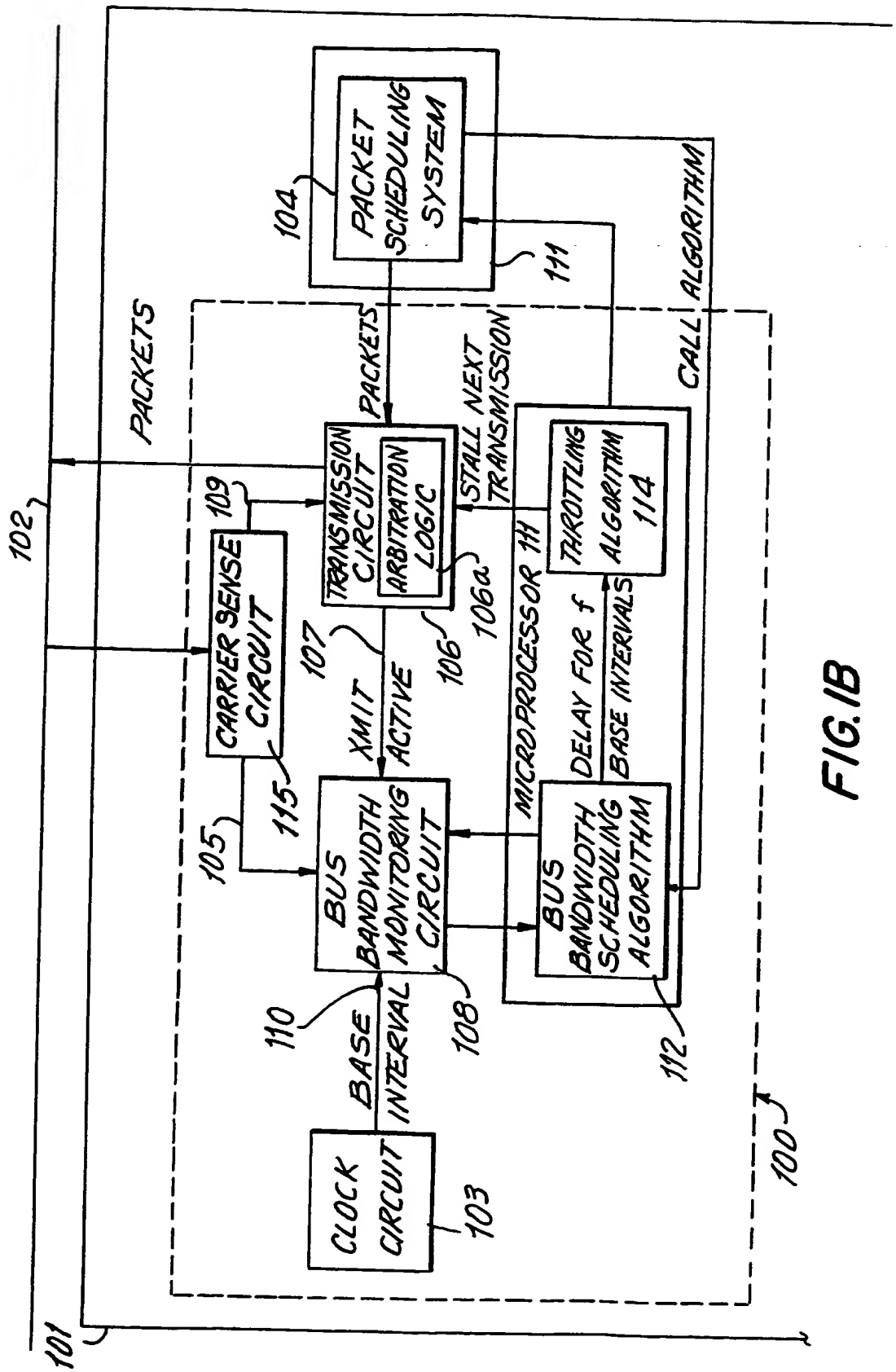

PACKET SCHEDULING SYSTEM
104
111
ALGORITHM
CALL
PACKETS
PACKETS
102
109
TRANSMISSION CIRCUIT
ARBITRATION LOGIC
STALL NEXT TRANSMISSION
CARRIER SENSE CIRCUIT
107
XMIT ACTIVE
106
106a
MICROPROCESSOR 111
THROTTLING ALGORITHM 114
105
115
BUS BANDWIDTH MONITORING CIRCUIT
DELAY FOR f BASE INTERVALS
BUS BANDWIDTH SCHEDULING ALGORITHM
110
BASE INTERVAL
108
112
CLOCK CIRCUIT
103
100
101

FIG. 1B

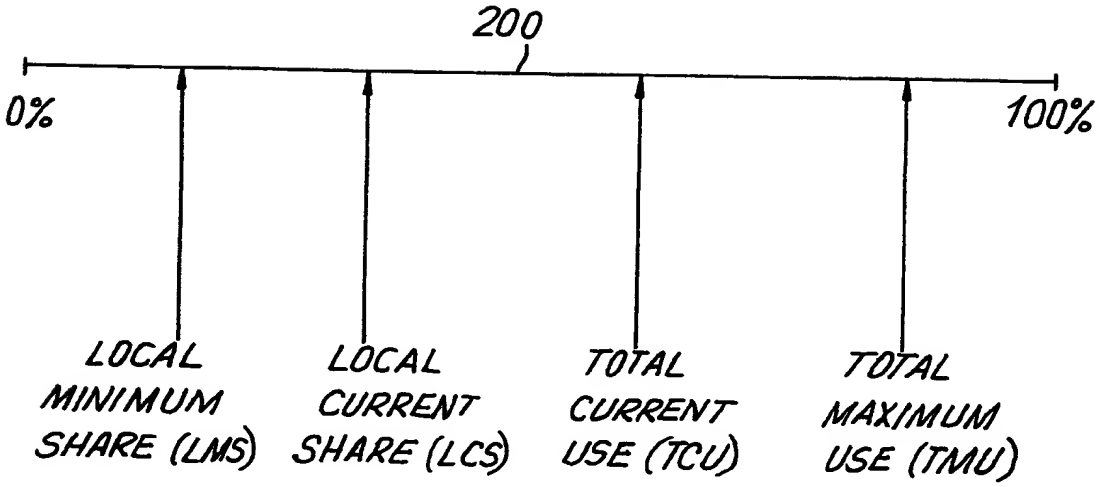

200
0%
100%
LOCAL MINIMUM SHARE (LMS)
LOCAL CURRENT SHARE (LCS)
TOTAL CURRENT USE (TCU)
TOTAL MAXIMUM USE (TMU)

FIG. 2

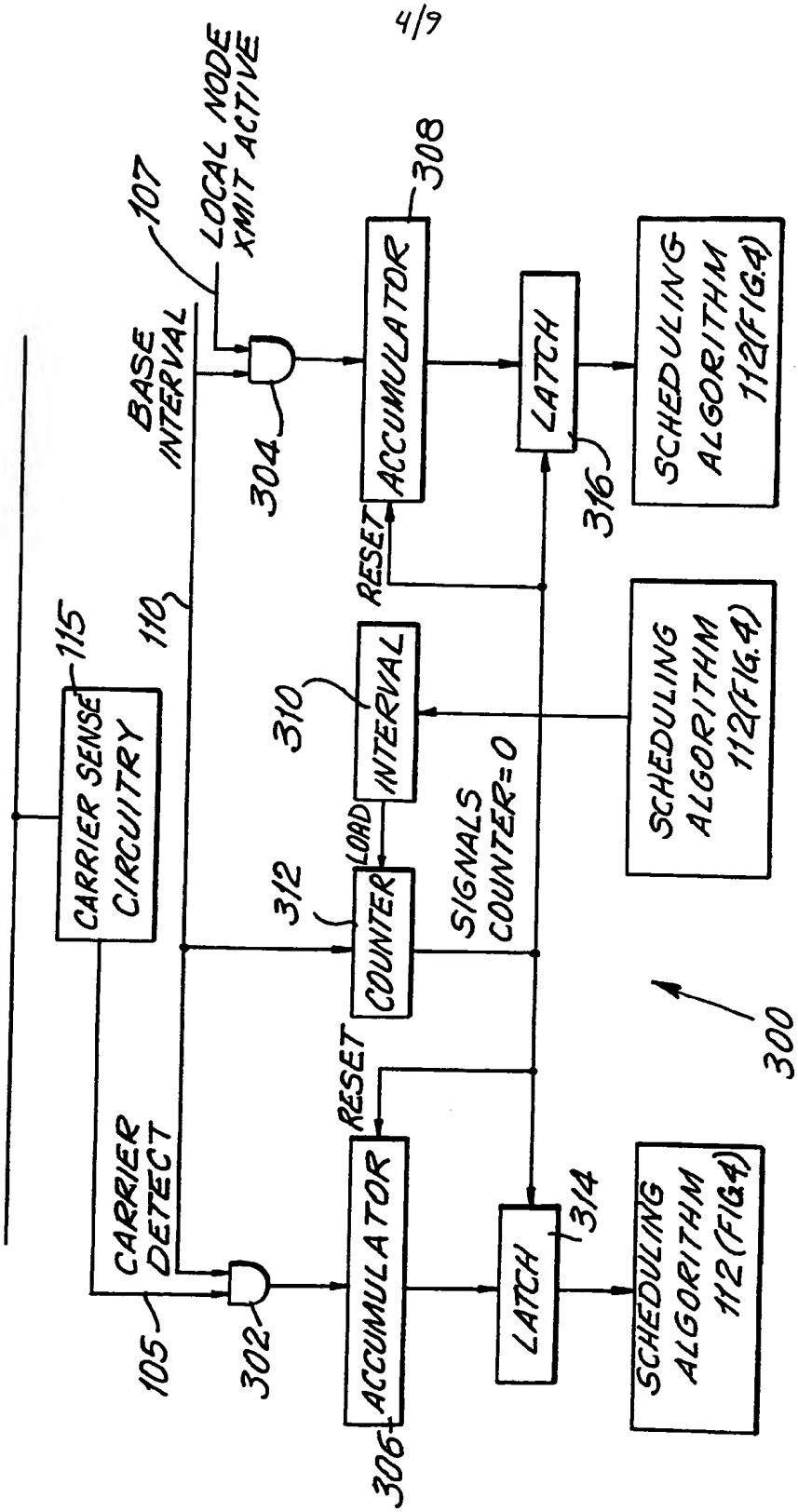

LOCAL NODE XMIT ACTIVE
107
BASE INTERVAL
308
304
ACCUMULATOR
LATCH
SCHEDULING ALGORITHM 112(FIG4)
316
RESET
110
115
CARRIER SENSE CIRCUITRY
310
INTERVAL
LOAD
SCHEDULING ALGORITHM 112(FIG.4)
312
COUNTER
SIGNALS COUNTER=0
RESET
CARRIER DETECT
105
302
ACCUMULATOR
306
LATCH
314
SCHEDULING ALGORITHM 112 (FIG4)
300

FIG.3

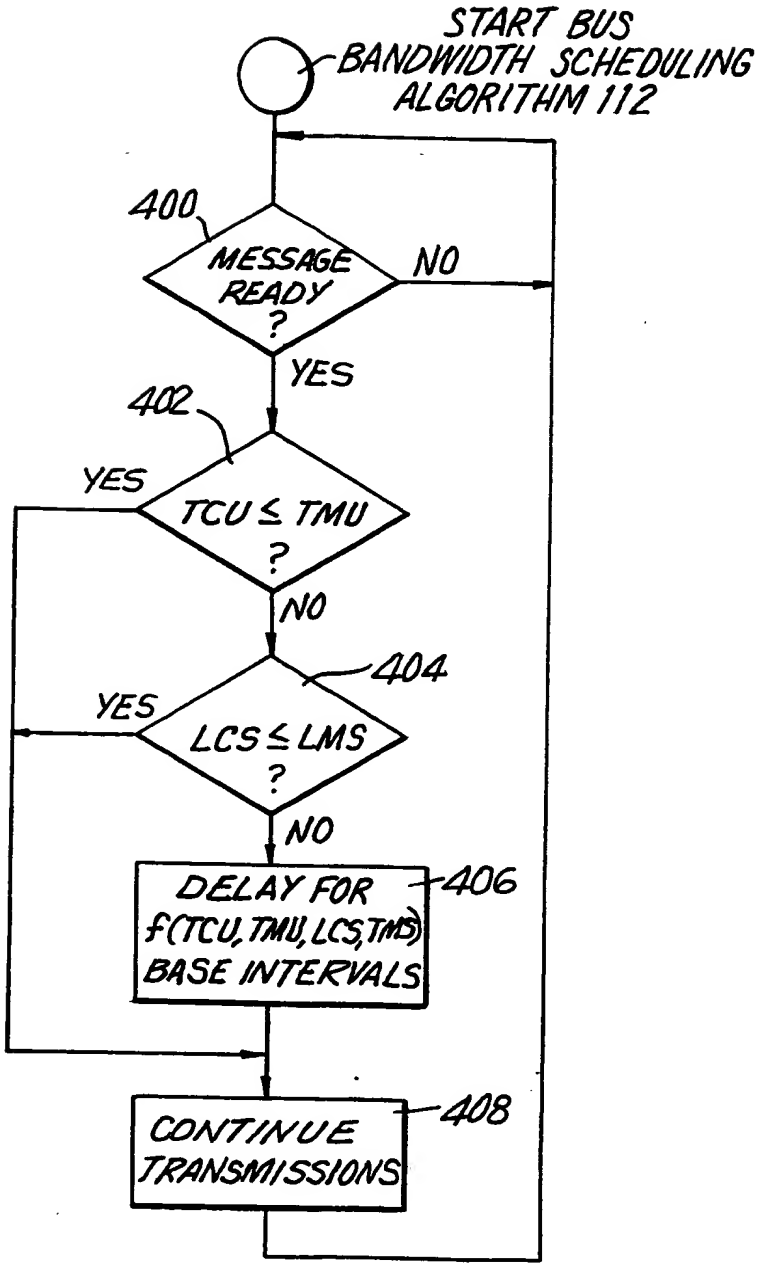

START BUS BANDWIDTH SCHEDULING ALGORITHM 112
400
MESSAGE READY ?
NO
YES
402
YES
TCU ≤ TMU ?
NO
404
YES
LCS ≤ LMS ?
NO
DELAY FOR f(TCU, TMU, LCS, TMS) BASE INTERVALS
406
CONTINUE TRANSMISSIONS
408

FIG. 4

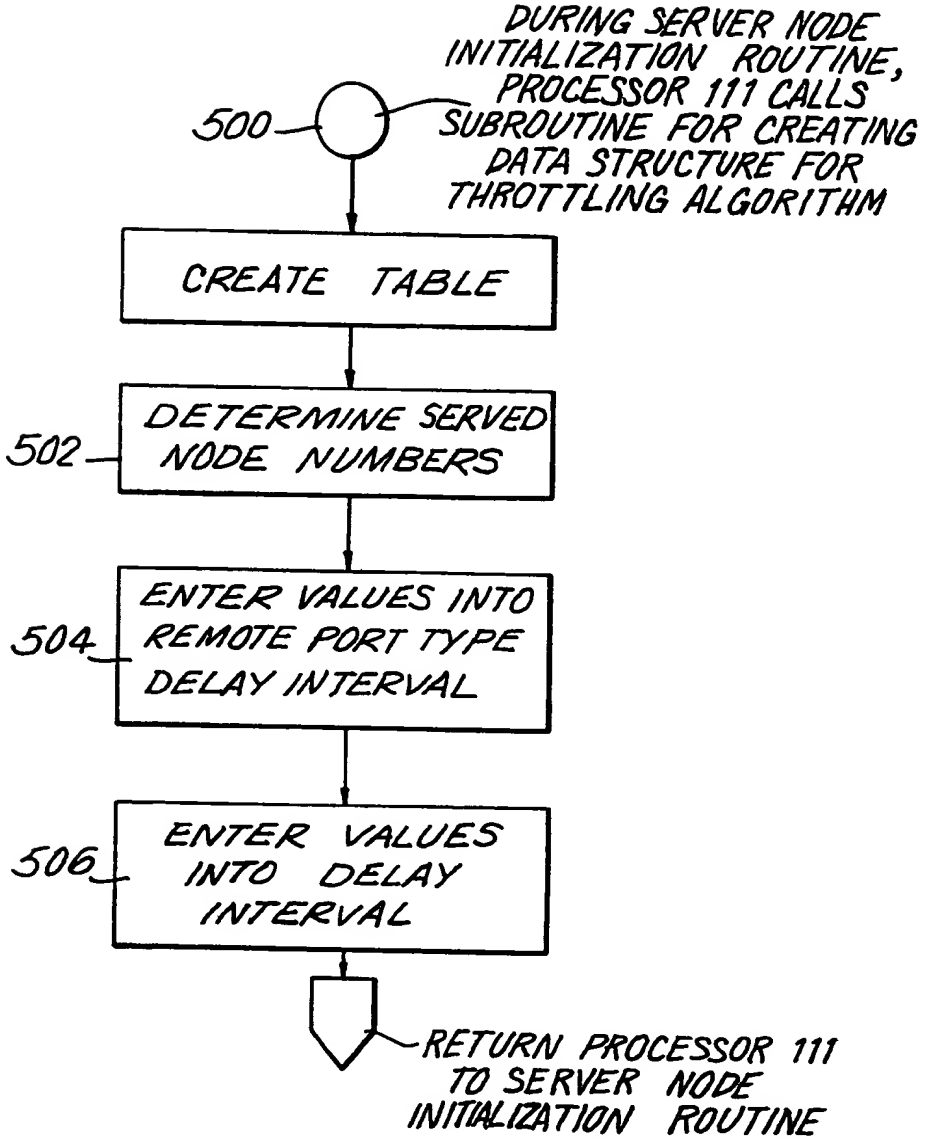

DURING SERVER NODE INITIALIZATION ROUTINE, PROCESSOR 111 CALLS SUBROUTINE FOR CREATING DATA STRUCTURE FOR THROTTLING ALGORITHM
500
CREATE TABLE
DETERMINE SERVED NODE NUMBERS
502
ENTER VALUES INTO REMOTE PORT TYPE DELAY INTERVAL
504
ENTER VALUES INTO DELAY INTERVAL
506
RETURN PROCESSOR 111 TO SERVER NODE INITIALIZATION ROUTINE

FIG.5A

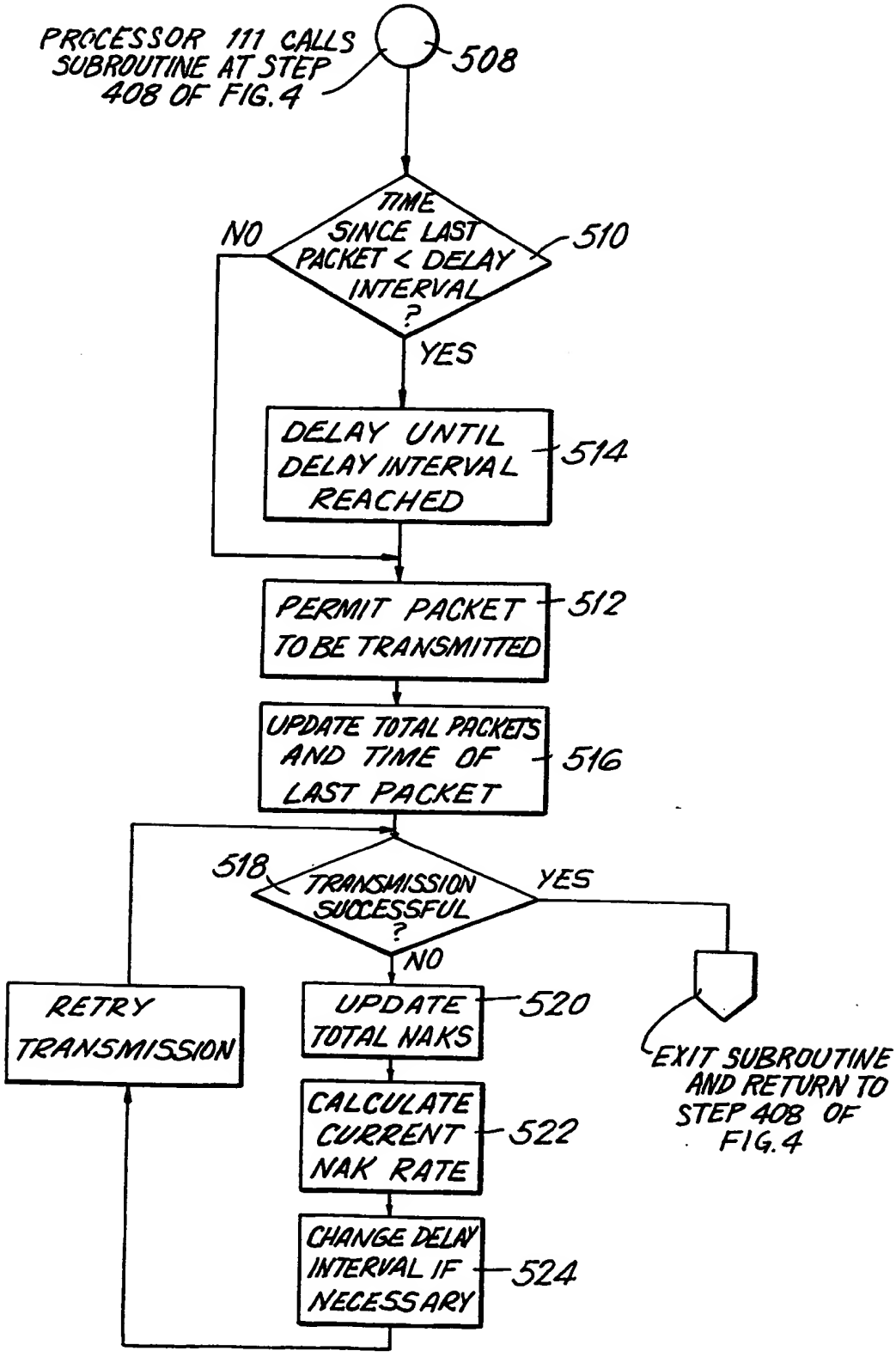

PROCESSOR 111 CALLS SUBROUTINE AT STEP 408 OF FIG. 4
508
TIME SINCE LAST PACKET < DELAY INTERVAL ?
510
NO
YES
DELAY UNTIL DELAY INTERVAL REACHED
514
PERMIT PACKET TO BE TRANSMITTED
512
UPDATE TOTAL PACKETS AND TIME OF LAST PACKET
516
518
TRANSMISSION SUCCESSFUL ?
YES
NO
RETRY TRANSMISSION
UPDATE TOTAL NAKS
520
CALCULATE CURRENT NAK RATE
522
CHANGE DELAY INTERVAL IF NECESSARY
524
EXIT SUBROUTINE AND RETURN TO STEP 408 OF FIG. 4

FIG. 5B

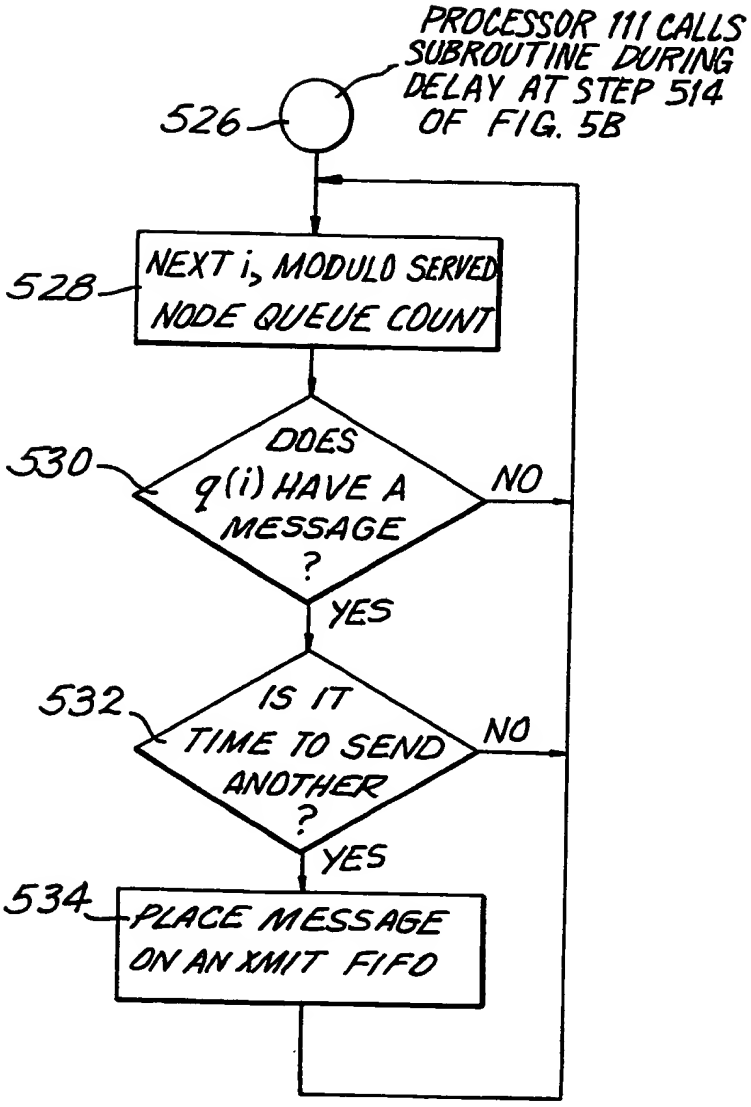
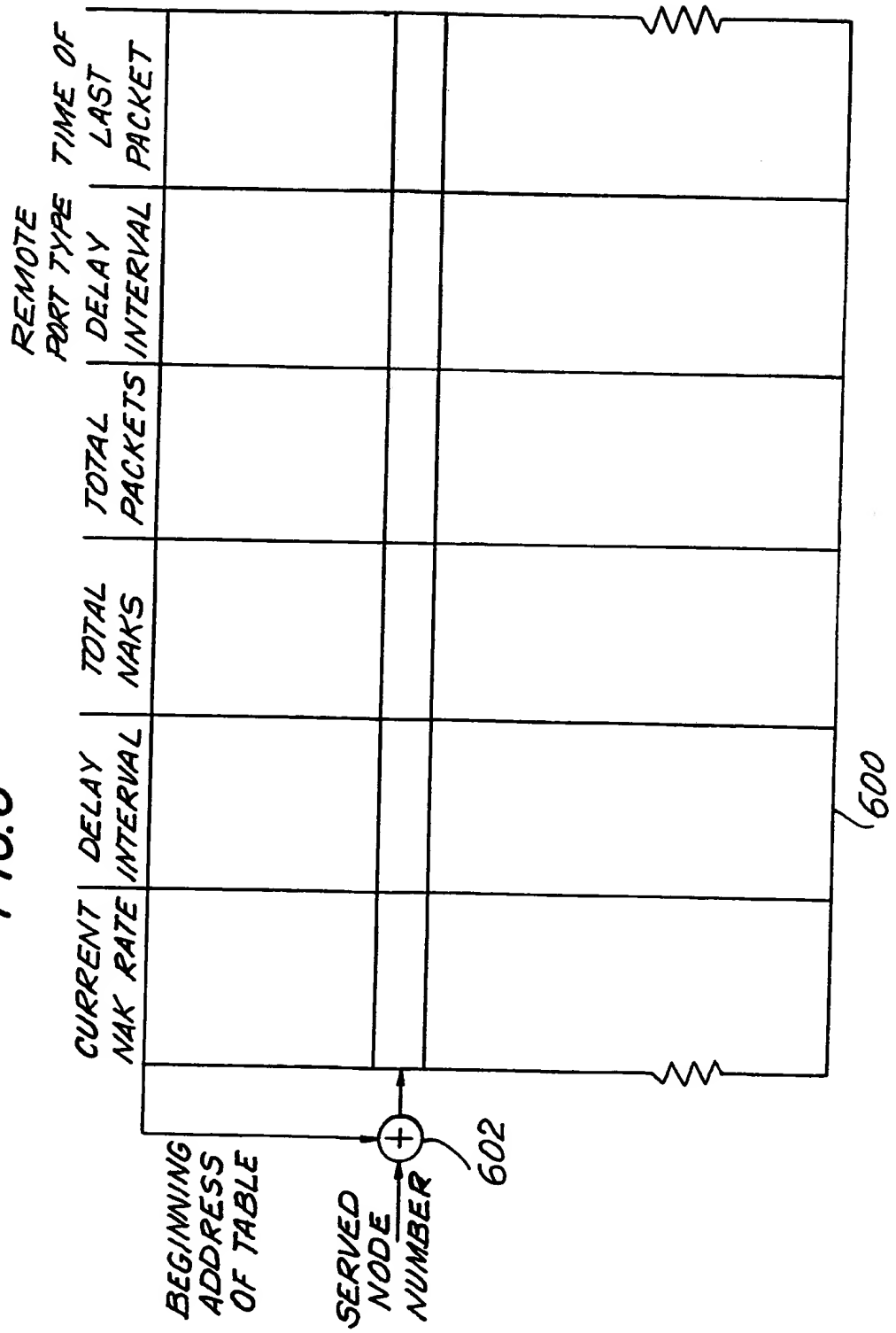

PROCESSOR 111 CALLS SUBROUTINE DURING DELAY AT STEP 514 OF FIG. 5B
526
NEXT i, MODULO SERVED NODE QUEUE COUNT
528
DOES q(i) HAVE A MESSAGE ?
530
NO
YES
IS IT TIME TO SEND ANOTHER ?
532
NO
YES
PLACE MESSAGE ON AN XMIT FIFO
534

FIG. 5C

FIG.6


CURRENT NAK RATE
DELAY INTERVAL
TOTAL NAKS
TOTAL PACKETS
REMOTE PORT TYPE
DELAY INTERVAL
TIME OF LAST PACKET
600
602
BEGINNING ADDRESS OF TABLE
SERVED NODE NUMBER

INTERNATIONAL SEARCH REPORT

International Application No PCT/US 91/01310

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all)[6]

According to International Patent Classification (IPC) or to both National Classification and IPC

Int.Cl. 5 H04L12/40 ; H04L29/06

II. FIELDS SEARCHED

Minimum Documentation Searched[7]

Classification System	Classification Symbols
Int.Cl. 5	H04L

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched[8]

III. DOCUMENTS CONSIDERED TO BE RELEVANT[9]

Category °	Citation of Document,[11] with indication, where appropriate, of the relevant passages[12]	Relevant to Claim No.[13]
A	EP,A,241113 (LIMB) 14 October 1987 see column 3, line 43 - column 4, line 21 see column 4, line 50 - column 5, line 47	1, 4-13
A	DE,A,3613898 (SIEMENS) 29 October 1987 see column 3, line 52 - column 4, line 3 see column 4, line 35 - column 5, line 10 see column 7, lines 28 - 37 see column 7, lines 49 - 55	1, 4-13
A	INTERFACES IN COMPUTING. vol. 1, 1983, LAUSANNE CH pages 255 - 262; A.ARATO et al.: "A LOCAL AREA NETWORK ARCHITECTURE TAILORED TO LABORATORY ENVIRONMENTS" see paragraphs 3 - 4	1, 2, 8, 11, 12, 18, 19

-/--

° Special categories of cited documents :[10]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report
21 JUNE 1991	01.08.91
International Searching Authority	**Signature of Authorized Officer**
EUROPEAN PATENT OFFICE	MIKKELSEN C.

Form PCT/ISA/210 (second sheet) (January 1985)

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)

Category °	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claim No.
A	PATENT ABSTRACTS OF JAPAN vol. 9, no. 65 (E-304)(1788) 26 March 1985, & JP-A-59 204341 (MATSUSHITA) 19 November 1984, see the whole document	1, 4-13
A	1987 SYMPOSIUM ON THE SIMULATION OF COMPUTER NETWORKS August 1987, IEEE NEW YORK US pages 123 - 127; R.SIGNORILE et al.: "A STUDY OF A PRIORITY PROTOCOL FOR PC BASED LOCAL AREA NETWORKS SUPPORTINGLARGE FILE TRANSFERS" see paragraphs 2 - 3	1, 2, 8, 11, 12, 18, 19

ANNEX TO THE INTERNATIONAL SEARCH REPORT ON INTERNATIONAL PATENT APPLICATION NO.

US9101310
SA 45242

This annex lists the patent family members relating to the patent documents cited in the above-mentioned international search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information. 21/06/91

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP-A-241113	14-10-87	US-A- 4779267	18-10-88
DE-A-3613898	29-10-87	None	